Earth Syst. Sci. Data Discuss., 8, C18–C25, 2015
www.earth-syst-sci-data-discuss.net/8/C18/2015/

Open Access Earth System Science Data Discussions

Interactive comment on "A metadata template for ocean acidification data" *by* L.-Q. Jiang et al.

C. Chandler (Referee)

cchandler@whoi.edu

Received and published: 4 March 2015

GENERAL COMMENTS:

This reviewer agrees strongly with the authors that metadata (data set documentation) is essential for ensuring that a data set is accessible and reusable in the future. Furthermore, as noted by the authors, support for automated, machine discovery and interpretation of data will be greatly facilitated by structured metadata that follows conventions adopted by a community of practice and for which a template has been made available. Ocean acidification (OA) has emerged as an important research topic in marine biogeochemistry, involving in situ chemistry measurements, and observations of biological response in the field as well as laboratory conditions and mesocosm experiments, and the full complement of molecular biology studies. Such a vast range of measurement types demands rigorous data management best practices if the full

potential of the resultant data sets is to be realized. Therefore the topic of this paper, "A metadata template for ocean acidification data", is timely and the recommendations will be well received. The authors have considerable experience in the management of carbon cycle science data as evidenced by the recommendations discussed.

The rest of this review pertains to the template itself that is essentially an extension of previous recommendations for carbon cycle chemistry data sets.

STRENGTHS: The template covers the usual collection of metadata categories including the corresponding information relating to the 'who, what, where, when, and how questions' of a research data set, and the authors recommend adoption of international standards such as ISO 19115. Such metadata is essential, and the authors highlight the particular importance of sampling and analytical protocols and quality control documentation necessary to enable future users of the data to determine whether the data meet their research needs, e.g. 'fitness for purpose'.

WEAKNESSES: It is not clear who the intended audience is for this template, original investigators, data managers at thematic data repositories, research librarians and/or data archivists at large national or world data centers? It would be helpful to propose a suggested list of users of this template and describe the mechanism by which the authors expect such a template to be adopted by the international community.

While there is clear value to having a metadata template that serves to recommend a common set of guidelines based on emerging best practices, this reviewer encourages the authors to take this opportunity to include stronger recommendations for content guidelines. In particular, these concerns relate to the achievement of goal #3 (page 4, section 2.1), the goal of creating a template that can be adopted by the international community. See details below in the section on 'specific comments'.

SPECIFIC COMMENTS:

STRENGTHS: Table 3 on page 21, the 'Commonly used observation types', is a valu-

able set of observation modes. This list is a valuable contribution to the community. I encourage the authors to share it with other practitioners and then pursue registration with an appropriate system that has a functioning governance system (e.g. the NERC Vocabulary Server (NVS) hosted at the British Oceanographic Data Center; NVS2.0, http://vocab.nerc.ac.uk/; http://www.bodc.ac.uk/products/web_services/vocab/). This would provide a mechanism for terms to be assigned URIs such that the terms can be dereferenced online, and adopted by a wider community. Observation or data collection mode is a common way for researchers to search for data of interest, and it is therefore important to make that information available in a machine discoverable way. The authors have already contacted the Ocean Acidification International Coordination Center (OA-ICC); http://www.iaea.org/ocean-acidification/) and this reviewer strongly encourages them to follow up on that collaboration. If the authors agree, it would be of value to mention this, perhaps in the Acknowledgements, and as a way to recognize the work already being done by the OA-ICC (see reviewer comments from J.-P. Gattuso, ESSDD 8, C1–C3, 2015).

WEAKNESSES:

Most of the remaining suggestions are based on this reviewer's opinion that the template should make stronger recommendations for the way in which the variables in Table 1 and the 'Child elements' in Table 2 are reported. For example, this was done on page 11, lines 5-10, with the recommendation to use the WoRMS registry for species names. This reviewer agrees with the recognition that groups may already being using registered code lists, and should continue to have the freedom to do so. For maximum benefit, recommendations should be made for suggested content guidelines for more of the fields. In general, a strong recommendation should be made that wherever possible globally unique, persistent identifiers be used to identify the primary term in each metadata concept category, e.g. person, platform, instrument, etc. Much greater value can be realized from terms that are identified by and referenced as URIs (in addition to text string labels). For example,

the Pan-European SeaDataNet community has made considerable progress in this area, successfully using vocabulary lists available from NVS (e.g. 32OC ICES code from NVS, identified by URI: http://vocab.nerc.ac.uk/collection/C17/current/32OC). In many cases, the NVS vocabulary lists have also been translated into languages other than English, which makes them very useful for supporting international collaboration. When code lists are made available online and backed by a governance system that manages extension and deprecation of terms, they have the potential to become powerful tools to enable machine discovery, interpretation and interoperability. The full list of vocabularies already adopted by the SeaDataNet community is available: (http://seadatanet.maris2.nl/v_bodc_vocab_v2/welcome.asp).

The practice of using URIs to disambiguate terms in metadata is relatively new and this paper is a great opportunity to encourage wider adoption of this practice. It is important to make the recommendation, as the authors did in recommending use of WoRMS for species identification, such that data managers are encouraged to continuing using controlled vocabularies already in use or discover new term lists as they become available. Many of the recommendations below have already been widely adopted by data managers, institutional repository librarians and data archivists in the marine research community. Shared community practices have been promoted through several trans-national and global-scale networks, most notably: the IOC International Oceanographic Data and Information Exchange (IODE; iode.org); the Ocean Data Interoperability Platform (ODIP; odip.org); and the Marine Data Harmonization Interest Group of the Research Data Alliance (rd-alliance.org.org).

Some examples, with specific controlled vocabulary recommendations that could be added to the relevant description sections:

3.1.1 Key child elements of a variable (Details of child elements listed in Table 2) In biological response studies, it is necessary to report the life stage of the organism studied. This was also noted by reviewer Y. Yang (ESSDD-8-C4-2015). This recommendation could easily be added to section 3.1.1 'Key child elements of a variable', page 9, lines 8-

13, in the paragraph that measurements of biological response of a specific organism. The NVS2.0 S11 vocabulary ("S11: BODC parameter semantic model biological entity development stage terms" http://vocab.nerc.ac.uk/collection/S11/current/all/ would be a reasonable controlled vocabulary.

3.1.2 Additional child elements of a variable 1. variable names and definitions should come from a controlled vocabulary, the 'Climate and Forecast standard name' is an excellent suggestion as it is in wide use, but does not include many of the variables that would be reported for OA research results. Several NVS SeaDataNet Parameter Discovery and Usage vocabularies (P01 through P08, plus other P* series vocabularies) should be recommended as well: P08: http://vocab.nerc.ac.uk/collection/P08/current/all/ SeaDataNet Parameter Disciplines P03: http://vocab.nerc.ac.uk/collection/P03/current/all/ SeaDataNet Agreed Parameter Groups P01: http://vocab.nerc.ac.uk/collection/P01/current/all/ BODC Parameter Usage Vocabulary

2. page 9, line 26 through page 10, line 21: A recommendation should be made that where possible Sampling and Analyzing instrument should be identified with URIs that can be dereferenced (e.g. the NVS L05 and L22 vocabularies) L05: http://vocab.nerc.ac.uk/collection/L05/current/all/ SeaDataNet device categories L22: http://vocab.nerc.ac.uk/collection/L22/current/all/ SeaVoX Device Catalogue

3. page 11, lines 1-4, data quality flag: take the opportunity to make a recommendation to help get maximum benefit from use of the template. This reviewer recommends the "Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data" from the Intergovernmental Oceanographic Commission of UNESCO. 2013. Ocean Data Standards, Vol.3: Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data. (IOC Manuals and Guides, 54, Vol. 3.) 12 pp. (English.)(IOC/2013/MG/54-3) http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=10762. The quality flag schema is one of several from the Ocean Data Standards and Best

Practices Project (ODSBP; http://www.oceandatastandards.org/).

Section 3.2 Investigators Here the authors point out the value of using DOIs to identify publications. This is an opportunity to recommend the analogous practice of using person identifiers (e.g. ORCID, Researcher ID, etc.) to unambiguously define the investigator, and recommend using a controlled vocabulary for organizations as well. A controlled vocabulary for organization name to identify an investigator's institutional affiliation is not so well defined at this time, but the most promising thus far, the European Directory of Marine Organisations (EDMO) code list, is being promoted within SeaDataNet.

3.3 Temporal and spatial coverage For temporal coverage, ISO time standards should be used, and gazetteer terms in addition to standard WESN bounds; see Marine Regions standard list of marine georeferenced place names and areas (http://www.marineregions.org/downloads.php).

3.4 Platforms and sampling IDs The authors appear to conflate platform and deployment in this section. It is important to clearly identify both instances, in the case of field studies involving platform deployments, e.g. a research cruise which is the deployment of a research vessel (the platform).

1. for platform identifier, the authors recommend ICES code or NOAA ship code from a local code list. Populating the metadata with a mixture of ICES codes, recognized internationally, and local NOAA ship codes, will likely lead to confusion about the very content the code lists are intended to clarify, unless the terms are accompanied by URI's that clearly identify the namespace in which those terms are defined. Many organizations have developed local code lists over time, and several of these code lists have been adopted by a wider community. The ICES code list available as a well-ordered, controlled vocabulary (the C17 vocabulary from NVS is one example as described earlier).

2. deployment identifier: original expedition identifier, typically assigned by the operator

of the platform; when combined with the globally unique platform code (e.g. ICES code), makes a unique identifier string. The EXPOCODE is an example of a community identifier that was widely adopted during the WOCE years (beginning in the 1990s) and has continued to be used by the CLIVAR community and others. If EXPOCODE lists have been exposed openly, with associated URIs, then these could be valuable as well beyond the initial communities of practice.

3.5 Funding agencies and projects Including funding source information is a relatively recent practice, and the authors point out the significance of crediting the funding source for US funded research. Once again, this reviewer suggests going further and recommending that as a best practice, where possible (e.g. the information is available and known), the content include relevant award identifiers (accompanied by URIs from the funding source that can be dereferenced). While a related concept, the project name is often separate from the funding source, and should be clearly identified and reported in addition to the funding. In many cases, the project name can serve as a valuable search term aiding resource discovery.

3.6 Data citation, references, and supplementary information This is another opportunity to refer to community best practice. In this case, this reviewer suggests citing the FORCE11 Joint Declaration of Data Citation Principles (https://www.force11.org/datacitation).

Table 1. Some commonly used ocean acidification variables, their definitions, and recommended abbreviations.

Table 1 makes reference to the fact that variables are reported with units of measure. The units of measure are critical to include in metadata, and that point should be made more strongly and furthermore that units should be reported in accordance with the NIST International System of Units (SI).

TECHNICAL CORRECTIONS:

This reviewer found no technical corrections. The manuscript is well-written with content organized logically.

Interactive comment on Earth Syst. Sci. Data Discuss., 8, 1, 2015.